Developmental regulation of canonical and small ORF translation from mRNAs

Article (Published Version)

Patraquim et al. Genome Biology (2020) 21:128
https://doi.org/10.1186/s13059-020-02011-5

Genome Biology

RESEARCH Open Access

Developmental regulation of canonical and small ORF translation from mRNAs

Pedro Patraquim[1†], Muhammad Ali Shahzad Mumtaz[2†], José Ignacio Pueyo[2], Julie Louise Aspden[3] and Juan-Pablo Couso[1,2*]

* Correspondence: jpcou@upo.es
[†]Pedro Patraquim and Muhammad Ali Shahzad Mumtaz contributed equally to this work.
[1]Present Address: Centro Andaluz de Biologia del Desarrollo, CSIC-UPO, Seville, Spain
[2]Brighton and Sussex Medical School, Brighton, East Sussex, UK
Full list of author information is available at the end of the article

Abstract

Background: Ribosomal profiling has revealed the translation of thousands of sequences outside annotated protein-coding genes, including small open reading frames of less than 100 codons, and the translational regulation of many genes. Here we present an improved version of Poly-Ribo-Seq and apply it to *Drosophila melanogaster* embryos to extend the catalog of in vivo translated small ORFs, and to reveal the translational regulation of both small and canonical ORFs from mRNAs across embryogenesis.

Results: We obtain highly correlated samples across five embryonic stages, with nearly 500 million putative ribosomal footprints mapped to mRNAs, and compare them to existing Ribo-Seq and proteomic data. Our analysis reveals, for the first time in *Drosophila*, footprints mapping to codons in a phased pattern, the hallmark of productive translation. We propose a simple binomial probability metric to ascertain translation probability. Our results also reveal reproducible ribosomal binding apparently not resulting in productive translation. This non-productive ribosomal binding seems to be especially prevalent amongst upstream short ORFs located in the 5′ mRNA leaders, and amongst canonical ORFs during the activation of the zygotic translatome at the maternal-to zygotic transition.

Conclusions: We suggest that this non-productive ribosomal binding might be due to cis-regulatory ribosomal binding and to defective ribosomal scanning of ORFs outside periods of productive translation. Our results are compatible with the main function of upstream short ORFs being to buffer the translation of canonical canonical ORFs; and show that, in general, small ORFs in mRNAs display markers compatible with an evolutionary transitory state towards full coding function.

Keywords: Ribosomal profiling, Poly-Ribo-Seq, Ribosomal binding, Non-canonical translation, Regulation of translation, smORFs, sORFs, uORFs, Maternal to zygotic transition

Background

Ribosomal profiling, the genome-wide sequencing of ribosome-protected RNA fragments (Ribo-Seq), has been increasing our understanding of a crucial event in all living genomes: protein translation [1, 2]. Ribo-Seq results show that there exists a noticeable

level of non-canonical translation in eukaryotes. This can arise from non-AUG codons, from polycistronic transcripts, and from unannotated small open reading frames of less than 100 codons (smORFs or sORFs) [3–5].

At the same time, the comparison of RNA-Seq and Ribo-Seq for the same biological samples allows the transcriptome-wide study of an important but often forgotten regulatory process in genome function: the regulation of translation. Typically, expression of an mRNA transcript is equated with the automatic translation of any encoded canonical ORF (of more than 100 codons). However, translation is not automatic, but is regulated, often with great relevance for organismal function, both under normal and altered conditions [6–10]. Translation regulation is also a key aspect of development in all animals [11] and has been studied extensively in *Drosophila* [12–14]. *Drosophila* embryogenesis is a highly coordinated and complex process that is completed in a time span of just 24 h [15]. During the first 2 h after egg laying (AEL), there is absence of transcription from the zygotic genome and the key developmental processes, such as establishment of the primary antero-posterior and dorso-ventral axes, are controlled purely through the translational regulation of maternal mRNA previously laid down in the egg [16, 17]. After this initial period, the embryo undergoes a "maternal to zygotic transition," whereby the transcription and translation of the zygotic genome takes over the maternal products, a process also found in nematodes, echinoderms, and vertebrates [13, 18–20]. Nonetheless, the impact of translational regulation at the genome-wide scale on the whole of *Drosophila* embryogenesis has not yet been revealed.

Ribo-Seq results regarding non-canonical and regulated translation have been the subject of debate. While it has become accepted that both processes may occur more extensively than previously thought, there is no consensus on the actual fraction of smORFs and non-canonical ORFs whose translation is shown by Ribo-Seq [21–25]. The Ribo-Seq debate centers on the asymmetry between these numbers and other translational evidence, and on the interpretation of the Ribo-Seq results themselves. The most widely used counterpart of Ribo-Seq is proteomics, but the numbers of proteins and peptides detected by proteomics consistently fall short of those detected by Ribo-Seq, especially regarding non-canonical translation. For example, the most thorough proteomics study to date covering the whole *Drosophila melanogaster* life-cycle has detected less than 40% of all unique canonical proteins [26]. This number is further reduced to 30% of annotated smORF polypeptides, while we have previously reported that 80% of canonical and small ORFs show clear Ribo-Seq evidence of translation in a single embryonic cell line [23]. However, Ribo-Seq detects ribosomal binding, not actual peptide production. There is not a universally agreed Ribo-Seq metric unequivocally identifying productive, biologically relevant translation, as opposed to other processes such as low-level background translation, ribosomal scanning and nonsense-mediated-decay surveillance, or stochastic ribosomal binding. Bioinformatically, it is accepted that ribosomal binding above a certain level, and especially, binding showing trinucleotide periodicity in phase with codon triplets (phasing or framing), indicates translation of an ORF [1, 2]. A biochemical approach is to introduce modifications to the ribosomal-RNA purification, to ensure that only ribosomes engaged in productive translation are selected. For example, Ribo-Seq of polysomes (RNAs bound by several ribosomes), given that the sequential translation of polyadenylated, capped and circularized mRNAs by several ribosomes is a supramolecular feature of productive translation,

excludes single ribosomes (which could be involved in low-level translation but also in other activities) [23, 27]. We have called this latter approach Poly-Ribo-Seq [23].

Here we present an in vivo Poly-Ribo-Seq study covering a time-course of *Drosophila melanogaster* embryogenesis. We have both improved our experimental Poly-Ribo-Seq and the subsequent data analysis pipeline, to obtain unprecedented levels of Ribo-Seq efficiency (reads mapped to ORFs) and quality, including codon framing as the hallmark of productive, biologically meaningful translation. Thus, we can ascertain translation and its regulation in vivo and across development for both canonical and non-canonical ORFs. We detect the translation of thousands of non-annotated ORFs and identify hundreds of mRNAs whose translation is highly regulated during embryogenesis. However, our results also reveal reproducible ribosomal binding not resulting in productive translation. This non-productive ribosomal binding seems to be especially prevalent amongst upstream short ORFs located in the 5′ mRNA leaders, and amongst canonical ORFs during the activation of the zygotic translatome at the maternal to zygotic transition. We suggest that this type of ribosomal binding might be due to either cis-regulatory ribosomal activity, or to defective ribosomal scanning of ORFs outside periods of productive translation.

Results

The method and overall data

Since Poly-Ribo-Seq requires even larger amounts of starting material than Ribo-Seq due to polysome fractionation, the *Drosophila S2* cell line was an excellent tool for the development of the technique. However, the S2 cell line is derived from just one type of tissue (macrophage-like) from late stage *Drosophila melanogaster* embryos [28] and may not be the ideal system to obtain a comprehensive picture of *Drosophila* translation. Only 60% of both canonical genes and uORFs, and only about a third of annotated smORFs (hereafter referred as short CDSs, or sCDSs [29]) appear transcribed in this cell line [23]. Therefore, we applied Poly-Ribo-Seq to *D. melanogaster* embryos in order to extend the catalog of in vivo translated smORFs and study translation across three distinct developmental time windows of *Drosophila* embryogenesis (Fig. 1a). The original protocol from Aspden et al. [23] was adapted and improved in terms of biochemical rRNA depletion and bioinformatics mapping of ribosomal footprints (FPs) to ORFs (see "Materials and methods"), and we obtained higher yield (percentage of non-rRNA reads, Table S1), purity (percentage of ORF-mapping reads, Fig. 1b), and reproducibility (Fig. 1c and Table S2). This way, we detected rare but reproducible translation events in non-annotated genes and obtained a comprehensive picture of translation across development.

In our experimental design, we divided *Drosophila* embryogenesis into three temporal 8-h windows: early (0–8 h after egg laying, or 0–8 h AEL), mid (8–16 h AEL), and late embryogenesis (16–24 h AEL) (Fig. 1a). The key developmental processes occurring during these periods of embryogenesis are described elsewhere [15]. Briefly, early embryogenesis (0–8 h AEL) is characterized by the maternal to zygotic transition, followed by gastrulation, germ band formation, and ectodermal and mesodermal determination and pattern formation, including the determination of the CNS. The 8–16 h AEL of development (mid-embryogenesis) cover the process of tissue formation and

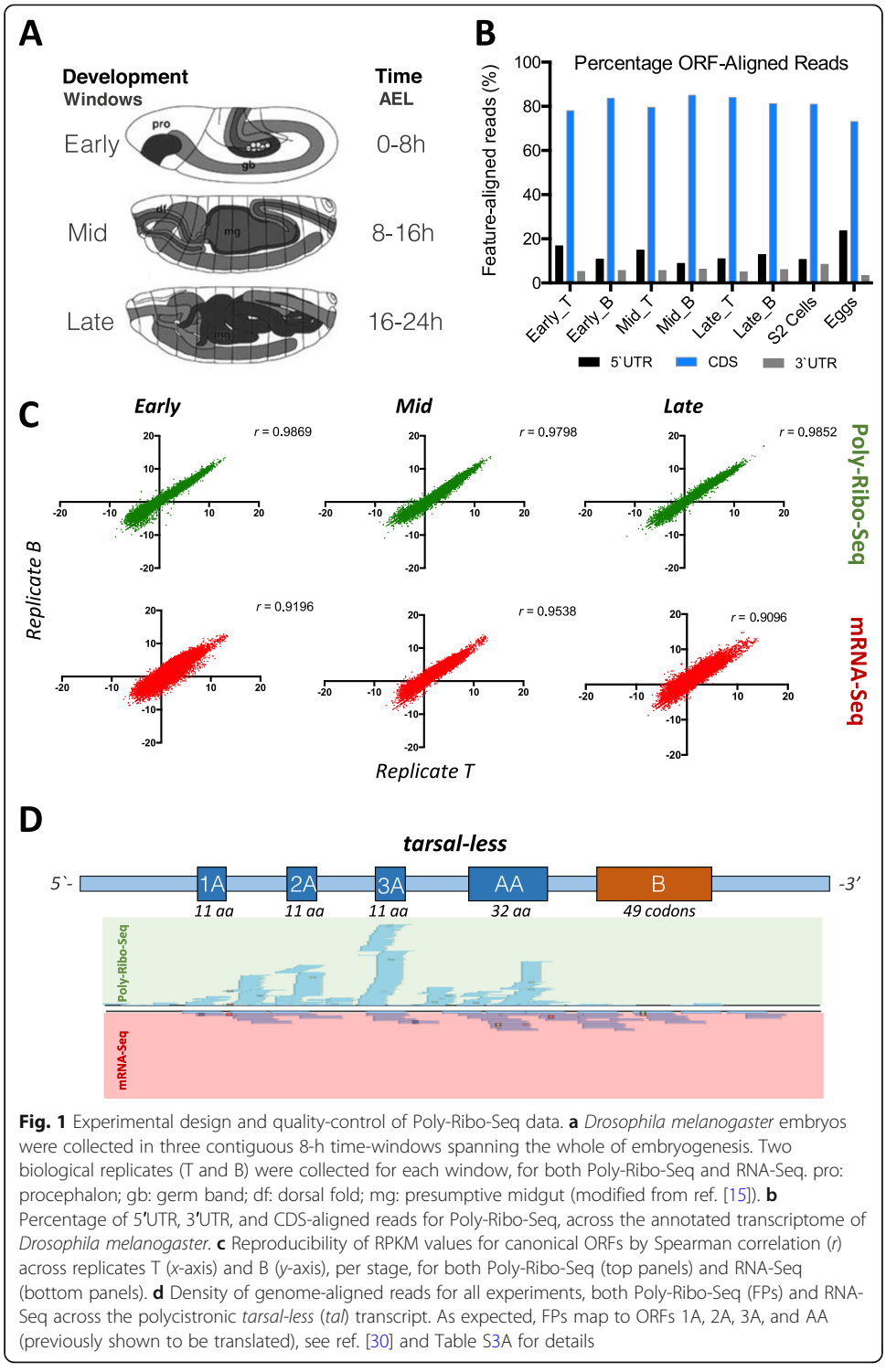

Fig. 1 Experimental design and quality-control of Poly-Ribo-Seq data. **a** *Drosophila melanogaster* embryos were collected in three contiguous 8-h time-windows spanning the whole of embryogenesis. Two biological replicates (T and B) were collected for each window, for both Poly-Ribo-Seq and RNA-Seq. pro: procephalon; gb: germ band; df: dorsal fold; mg: presumptive midgut (modified from ref. [15]). **b** Percentage of 5′UTR, 3′UTR, and CDS-aligned reads for Poly-Ribo-Seq, across the annotated transcriptome of *Drosophila melanogaster*. **c** Reproducibility of RPKM values for canonical ORFs by Spearman correlation (*r*) across replicates T (*x*-axis) and B (*y*-axis), per stage, for both Poly-Ribo-Seq (top panels) and RNA-Seq (bottom panels). **d** Density of genome-aligned reads for all experiments, both Poly-Ribo-Seq (FPs) and RNA-Seq across the polycistronic *tarsal-less* (*tal*) transcript. As expected, FPs map to ORFs 1A, 2A, 3A, and AA (previously shown to be translated), see ref. [30] and Table S3A for details

organogenesis (such as presumptive gonad and gut assembly) including multi-organ processes such as formation of segments and the feeding cavity. The differentiation of neurons and muscles also occurs during this time. Late embryogenesis (16–24 h AEL) is characterized by final differentiation and the start of physiological processes such as epidermal and cuticle differentiation [31] (with opening of tracheae to respiration) and

digestive system differentiation (with yolk digestion and uric acid production) [32, 33], plus the connection and fine-tuning of neural, sensory, and muscle structures, eventually resulting in coordinated embryo movements and larval hatching [34, 35].

We collected two biological replicas (called T and B) from each embryonic stage, extracting simultaneously total RNA and ribosomal footprints. Sequencing and analysis of these samples was used to determine transcription and translation levels, which in turn were used to reveal the extent and quality of translational regulation during *Drosophila* embryogenesis at a genomic scale. We studied several classes of ORFs as defined in Couso and Patraquim [29], altogether numbering 40,852 ORFs (see "Materials and methods"): 21,118 canonical ORFs of more than 100 codons in polyadenylated mRNAs; 862 ORFs from annotated coding sequences of 100 codons or less in an otherwise canonical mRNA (short CDSs); and 18,872 upstream ORFs (uORFs) in the 5′ leaders of canonical mRNAs. We studied only AUG-START ORFs and also discarded fully overlapping uORFs, in order to ensure the accuracy of our translation assessments. Similarly, the translation of ORFs in putative long-non-coding RNAs (lncORFs) will be described in a forthcoming study due to its special characteristics.

As the Ribo-Seq protocol typically results in a composite library with a majority of ribosomal rRNA reads, and a minority of ribosomal footprints (FPs), we sequenced an exhaustive amount, totalling 1.2 billion Ribo-Seq reads, yielding some 345 million ribosomal footprints (FPs) mapped to the genome (Table S1) with ~ 80% FPs aligned to annotated CDS, ~ 16% to 5′ leaders and 6% to 3′UTRs (Figs. 1b and 2a), providing a 300× coverage of our 40,852-strong ORF-ome. The data shows high reproducibility, with high Spearman correlation ($r = 0.9$) of RPKMs (number of sequenced reads per ORF length in kilobase, per million reads) for each ORF amongst the two replicas (Fig. 1c) for both RNA-Seq ($RPKM^{RNA}$) and Ribo-Seq ($RPKM^{FP}$) samples. Only reads of 26 to 36 nucleotides (nt) were counted for $RPKM^{FP}$, as this length distribution corresponded to 97% of the reads obtained in all cases and is within the size range previously described for ribosomal FPs [1] (Additional file 1: Fig. S1A). At first glance, our data revealed translation of non-canonical genes, as for the polycistronic smORF gene *tarsal-less*, where the FPs mapped to the ORFs experimentally proven to be translated [30] (Fig. 1d and Table S3A).

Framing: an indicator of productive translation

For our bioinformatic analysis (see "Materials and methods" and Fig. 2a), we use $RPKM^{RNA}>1$ to indicate transcription of an ORF (in either of our two replicas). To determine translation, we use two combined filters. First, a quantitative filter: a minimal amount of reproducible ribosomal binding must be detected to indicate signal above error and background; this "ribosomal bound" state was achieved by obtaining $RPKM^{FP}>1$ in both T and B replicas in a given stage. Second, a qualitative filter, the nature of that binding: the ribosomal movement across an ORF during translation generates tri-nucleotide phasing of ribosomal protected footprints, or framing (Fig. 2b), in contrast with the reads derived from RNA-Seq, which accumulate evenly across all nucleotides in mRNA (Additional file 1: Fig. S1B).

Previous ribosomal profiling studies of *Drosophila* had not been able to observe strong framing on a genome-wide scale, and thus translation has been proposed on the

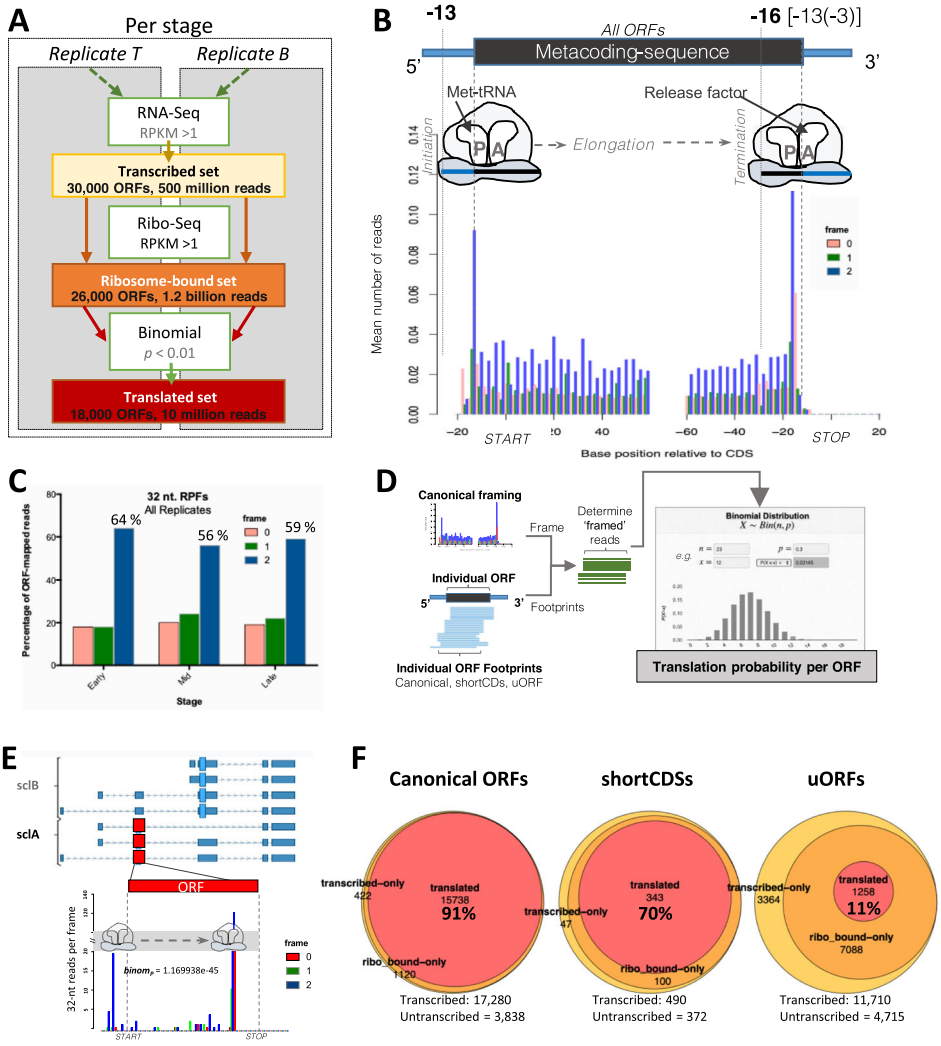

Fig. 2 Detection of translation in annotated and non-annotated ORFs. **a** Pipeline used in this study. An ORF was considered translated, in each stage, if it showed transcription (RPKM > 1 in one of two RNA-Seq replicates), reproducible ribosomal binding (RPKM > 1 in both Poly-Ribo-Seq replicates) and a significant framing of FPs (*p* value < 0.01 in the binomial test for framed-reads). **b** Metagene plot for codon framing of 32-nt ribosome footprints across all ORFs analyzed in this study, showing three-nucleotide periodicity (framing) in the third position of each codon (frame 2, blue). Numbers on top denote the distances between the 5′-end of ribosomal footprints and START/STOP codons. **c** The observed framing for 32-nt ribosome footprints is consistent across embryonic stages (combined replicates per stage). **d** Logic of per-ORF analysis of framing probabilities for each ORF at a given stage using the binomial test. **e** Density of 32-nt genome-aligned Poly-Ribo-Seq reads for all experiments, per frame, and corresponding translation probability for the *scl-A* ORF, as measured by the binomial test (see Table S3B for details). **f** Numbers of developmentally transcribed, ribosome-bound, and translated canonical ORFs (annotated ORFs with length > 100 codons), short CDSs (annotated ORFs ≤ 100 codons), and uORFs

basis of RPKM and other quantitative metrics, such as coverage [18, 23, 24, 36]. However, our improved Poly-Ribo-Seq method, coupled with an improved bioinformatics analysis of the data (see "Materials and methods"), revealed framing for the first time in *Drosophila* (Fig. 2b,c). Like all Ribo-Seq experiments, our data revealed a distribution of ribosomal footprint lengths (Additional file 1: Fig. S1A). We focused on the most represented read length (32 nucleotides), which also shows the clearest framing in all samples (56–64% reads in-frame; Fig. 2c). Different formulae and metrics have been

previously used to quantify framing per ORF and provide a cut-off [1, 21], reviewed in [5, 37], which was usually obtained by sampling a small number of bona-fide translated ORFs and extrapolating the results. While these methods can identify translation, they can be complex and difficult to apply universally to all Ribo-Seq datasets, due to differences in sampling and protocols utilized.

Here, we have opted for a simpler approach that is easy to extend to past and future datasets. We treat framing as a simple coin-toss problem, which offers clear statistical probabilities, even with a low number of reads. As each footprint is the result of an independent RNA-ribosome molecular interaction, each resulting sequencing read has in principle a 1/3 probability of aligning to the right frame in an ORF, and 2/3 of aligning to the wrong frame (Fig. 2d). The probability of having x successful outcomes (reads aligned to the ORF frame) following n independent trials (number of reads mapping to the ORF) follows a classical binomial distribution, which can be consulted using widely available tables [38] and computer programs. We take the observed binomial probability as the probability for a given ORF of being translated, and we consider it as indicative of translation when $p < 0.01$ (i.e., the probability of observing such framing of reads in the ORF by chance is less than one in a hundred; see "Materials and methods" for details and implementation). In principle, just five 32-nt FPs, all in frame in a given stage, are needed to pass a binomial test, but because of our $RPKM^{FP}>1$ filter (including FPs of 26 to 36 nt; see "Materials and methods" and Fig. S1), we noted that, out of 43,181 ORF translation events in all embryonic stages studied, only 21 events had less than 20 FPs, and no events had less than 10 FPs. As an example (Fig. 2e and Table S3B; see also Sup. Data for full sets of data), we show the non-canonical smORF gene *sarcolamban* (*scl*) whose translation has been determined experimentally [39]. Its two functional ORFs (*SclA* and *SclB*) achieve $RPKM^{FP}$ above 1 in both T and B replicas during mid and late embryogenesis. Most of its FP reads mapped to frame "2", but *SclB* only achieved a binomial p value below 0.01 during late embryogenesis. Thus, our data indicated that *SclA* is ribosomal-bound and translated at mid and late embryogenesis, but *SclB* is only productively translated during late embryogenesis, when *scl* function in regulating muscle contraction would have its maximum requirement (see ref. [39] and Fig. 1a). Note that framing p values do not indicate the amount of translation (which is indicated by the $RPKM^{FP}$ and its associated metric translational efficiency, TE, see below) but only the likelihood that the observed ribosomal binding leads to productive ORF translation. This binomial framing metric also ensures that the translation detected belongs to a given ORF, not to another putative ORF overlapping it.

Using these criteria, we observed that 98% of transcribed canonical ORFs, 90% of short CDSs, and 71% of uORFs were reproducibly bound by ribosomes at any point during development (showed $RPKM^{FP}>1$ in both replicates of the same stage), whereas the percentage of actually translated ORFs (ribosomal-bound ORFs also having framing p value < 0.01 in that stage) was 91% of Canonical, 70% of short CDSs, and 11% of transcribed uORFs (Fig. 2f).

Consistency of transcription and translation measurements across different types of genomic data

We further verified the robustness and reproducibility of our data and analysis by cross-comparisons across related gene-expression genomic data: RNA-Seq, Ribosomal

profiling, and quantitative proteomics. For RNA expression, we consolidated the modENCODE embryo RNA-Seq data [40], into our three 8-h-long developmental time windows. For all three developmental windows, the $RPKM^{RNA}$ of RNA-Seq samples was more highly correlated (in the region of $r = 0.9$) with a replica of their stage (either ours or modENCODE), than with samples of other stages (Figs. 1c, 3a, Table S2).

For ribosomal profiling, our $RPKM^{FP}$ data showed an even more marked trend to correlate preferentially first with their own stage replica ($r = 0.98–0.99$), and second with samples from adjacent time periods (Figs. 1c and 3a, Table S2). Our FPs and RNA-Seq from the same stage showed a high correlation of $r = 0.8–0.9$ (Fig. 1c, Sup Table 2). We also compared our FP data with the Kronja et al. 2014 [18] Ribo-Seq data from unfertilized *Drosophila* eggs, that was obtained after RNaseI digestion akin to our own protocol. As expected, the $RPKM^{FP}$ from unfertilized eggs (FP Maternal) showed the highest correlation with our $RPKM^{FP}$ samples from early embryogenesis (0–8 h AEL) (Spearman's $r \approx 0.8$) (Fig. 3a and Table S2). Our bioinformatic pipeline detected framing in Kronja's ORF-mapping reads of 29 nt (see "Materials and methods"). A total of 10,025 canonical ORFs appeared translated both in unfertilized eggs and early embryos, albeit showing increased translation efficiency in the latter (Fig. 3b). However, differences between the maternal and zygotic translatomes were also identified, since the Kronja data only assessed maternal RNA translation while our 0–8-h AEL sample detects both maternal and early zygotic translation (see "Regulation of canonical and sCDS translation during embryogenesis" below).

We also compared our embryo FP data with *Drosophila* embryo-derived S2 cell cultures [28], which in principle should provide synchronic data less subjected to experimental and developmental noise. We have re-analyzed the data from Aspden et al. [23], adding a new Poly-Ribo-Seq sample obtained here. S2 translation in general is highly correlated with early embryos (Fig. 3a), and the proportions of canonical ORFs ribosome-bound and translated are 98.0% and 65%, respectively. Despite this overall similarity, differences were also noted (see "Regulation of canonical and sCDS translation during embryogenesis" below).

For proteomics, we consolidated the results of extensive quantitative proteomics across embryogenesis [26] in our three 8-h developmental time windows ("Materials and methods"). This quantitative proteomics data showed an overall 0.45 correlation with our $RPKM^{FP}$ values (Fig. 3c). This correlation appears modest, but it is highly significant (at $p < 0.0001$), especially between $1<RPKM^{FP}<1024$, indicating that $RPKM^{FP}$ can be a good proxy of protein-production by canonical ORF translation at moderate to high levels. However, 40% of canonical ORFs detected as transcribed by RNA-Seq by us and modENCODE during embryogenesis are not detected by proteomics (Fig. 3d), while this number is reduced to 8% by our Poly-Ribo-Seq pipeline. The median $RPKM^{FP}$ value of the proteomics-detected proteins (32.11) is much higher than for those not detected (13.80), indicating that proteomics detection needs higher levels of protein translation, as noticed previously [21, 23]. Yet, many ORFs with experimentally verified translation, and with high Poly-Ribo-Seq metrics, are not detected by proteomics, including all *Hox* genes except the more widely expressed *Ubx* and *abd-A*. This "proteomics detectability" issue is not attributable to trypsination treatments, since we observed that 99% of canonical and 92% of non-canonical ORFs could produce K- and R-cleaved peptides of 7 to 24 amino acids (AA) suitable for mass-spec (MS) detection.

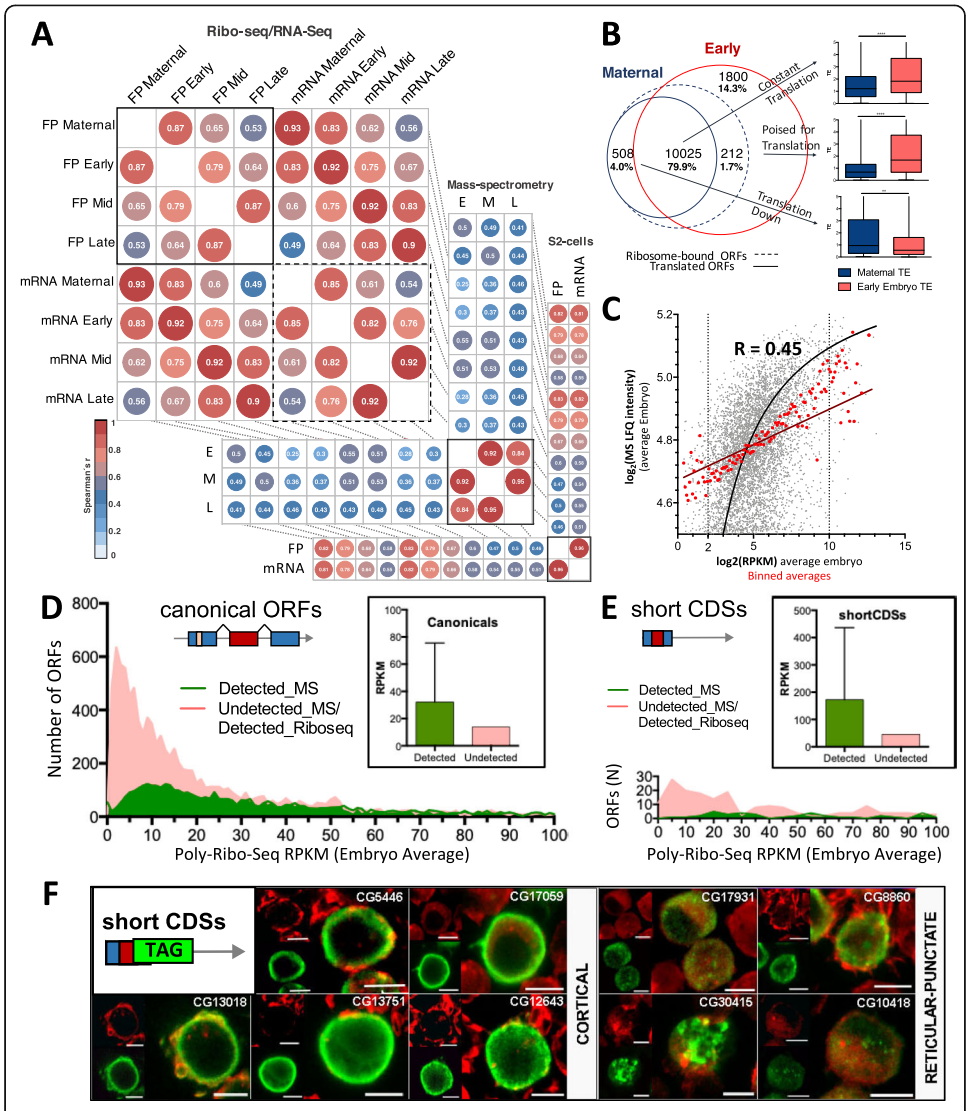

Fig. 3 Comparisons across genome-wide datasets. **a** Per-stage Spearman's correlations across Poly-Ribo-Seq ("FP"), RNA-Seq ("mRNA"), and mass-spectrometry datasets ("E"—early, "M"—mid, "L"—late stages correspondence in Casas-Vila et al. [26]—see "Materials and methods"). "Maternal" datasets correspond to concatenated mature oocyte and activated egg datasets from Kronja et al. [18]. S2-cells correspond to the Aspden et al. 2014 dataset plus new sequencing. Numbers denote Spearman's *rho*. **b** Numbers and translation fates of maternal ribosome-bound canonical ORFs in the maternal-to-zygotic transition. "Constant translation" denotes ORFs detected as translated by our pipeline in both maternal and early embryo datasets (see Fig. 2a). "Poised for translation" denotes ORFs showing maternal ribosome-binding which only appear translated in the early embryo. "Translation down" denotes ORFs with ribosome-binding and translation in the Maternal dataset, which do not appear translated in the early embryo dataset. **c** Overall correlation between average embryo Poly-Ribo-Seq RPKM and average embryo Mass-Spec imputed LFQ intensities (both datasets are $\log_2$-transformed). All detected ORFs are included in this analysis. Red points and line show the data binned in intervals of 0.1 RPKM. Note that the linear correlation is lost below RPKM = 1 ($\log_2 = 2$) and above RPKM = 1024 ($\log_2 = 10$). **d** Differential detectability of canonical ORFs between Poly-Ribo-Seq and mass spectrometry techniques. **e** Differential detectability of short CDSs between Poly-Ribo-Seq and mass spectrometry. **f** S2-tagging experiments of nine uncharacterized short CDSs showing translation in either cortical or reticular-punctate patterns (FLAG antibody: green, F-actin stained with phalloidin: red. Scale bar denotes 5 μm)

Hence, other yet unknown factors must influence the proteomics detection of proteins. In addition, small peptides of less than 50 amino acids suffer from enhanced degradation [41, 42], further hindering their detection by proteomics. Indeed, we observed that sCDSs are especially handicapped for MS detection, since the median $RPKM^{FP}$ of those detected is 172.1 (Fig. 3e). Finally, when comparing MS data with RNA-Seq and Poly-Ribo-Seq, and contrary to that observed between RNA and FP data (Fig. 3a), the highest correlation observed is amongst MS data from different embryonic stages. Altogether our analysis suggests that proteomics is an appropriate technique to detect constitutively and highly translated proteins, but not so useful for the detection of ORFs translated either moderately, or in a developmentally regulated manner.

smORFs also appeared harder to detect by Poly-Ribo-Seq than canonical ORFs, requiring a median $RPKM^{FP}$ of 45.8 (Fig. 3d,e, insets). As an independent indicator of translation, we have tagged a sample of smORFs from sCDSs and added to previous results [23, 43], for a total of 44 sCDSs. We compared the results of tagging with our Poly-Ribo-Seq pipeline, and with S2 cell proteomics data [44]. Out of the 44 tagged sCDSs, we observed that 43 were detected as translated by tagging, 27 by Poly-Ribo-Seq, and 14 by proteomics (Additional file 1: Fig. S2A). These results showed a coherent pattern, where the three techniques, based on very different detection technologies (Confocal microscopy vs. NextGenSeq vs. Mass-spec, respectively), show decreasing sensitivity thresholds: Tagging > Poly-Ribo-Seq > Proteomics, as also noted when comparing results from other species [45]. Thus, 11 of the 14 proteomics-detected sCDSs were also detected by Poly-Ribo-Seq, and 26 of the 27 detected by Poly-Ribo-Seq were also detected by Tagging. However, our expression of tagged peptides in S2 cells is based on non-endogenous high transcription and this might have pushed the peptide production of lowly expressed ORFs over the edge of detection (see "Materials and methods" and Aspden et al. 2014 [23]). Hence, our tagging results may indicate what "can be translated" rather than what actually "is". Accordingly, 14 of the 17 tag-positive sCDSs with no Poly-Ribo-Seq evidence in S2 cells were translated in other stages. The most abundant tagged peptide localizations (Fig. 3f) were (a) reticular-punctate in the cytoplasm, which may correspond to mitochondria [23], or else to mitochondria-associated ER producing the peptide and (b) cortical, which could indicate an association with the plasma membrane or its cytoskeletal cortex. These observations fit with the postulated tendency of sCDS peptides for membrane-associated localizations [23, 29].

Regulation of canonical and sCDS translation during embryogenesis

According to our data and modENCODE, 82% of genomic canonical ORFs and 57% of short CDSs were transcribed at any point during development (Fig. 2f). Of these, 9% and 30% respectively were never translated during embryogenesis, giving a first indication of the extent of translational regulation in *Drosophila melanogaster*.

Our data also showed that some 67% of translated canonical ORFs were detected constitutively (at all embryonic stages), whereas 33% were translated only at specific stages of development (Fig. 4a). Stage specificity seemed to be higher for smORFs, since only 53% of short CDSs and 17% of uORFs were detected constitutively across embryogenesis, whereas 47% and 83% respectively were translated only during some stages of

embryogenesis. However, this qualitative (translated vs. not translated) analysis reflected both transcriptional and translational regulation (since lack of mRNA precludes production of ribosomal footprints). To further quantify the extent and dynamics of purely translational regulation, we analyzed the changes in translational efficiency (TE) [2] across stages (TE indicates the ratio between FPs and RNA, calculated as $RPKM^{FP}$ / $RPKM^{RNA}$). This TE analysis revealed both on-off translation changes, and modulations of sustained translation, from one embryonic stage to the next. The average TE varied to some extent across development (Fig. 4b), but in general we observed fairly stable and correlated TE values across development (but see also "uORFs as translational regulators"). To identify which ORFs undergo statistically significant translational changes, we used the *Z*-ratio method (which basically indicates whether the variation of a given ORF TE is significantly outside the standard deviation of the total sample; see "Materials and methods") [46–48]. We also included in the analysis the Unfertilized Egg and S2 cells TE data (comparing them to early and late embryogenesis, respectively). We observed significant translational changes in all ORF classes, with the percentage of changed ORFs depending on the ORF class. About 12% of canonical proteins (1875 of 15,738 analyzed), 16% of short CDSs (55 of 343), and 19% of translated uORFs had significantly changed TE from any given stage to the next, indicating significant translational regulation (Fig. 4c). One of the most abundant changes was the translational upregulation of 529 canonical ORFs from early to mid embryogenesis (Fig. 4d). These ORFs were enriched for GO expression terms indicating a role in organogenesis (i.e., CNS, epidermis, and digestive system development) (Additional file 1: Fig. S2B), as befits the developmental processes underway (Fig. 1a).

The comparison of unfertilized eggs with 0–8 h AEL embryo data revealed the maternal to zygotic transition in the translatome. The degradation of maternal mRNAs after the zygotic to maternal transition from 2 h AEL, and their substitution by zygotically transcribed mRNAs has been reported, but a number of mRNAs do persist during early embryogenesis [49]. Our results suggest that some of these maternal mRNAs that persist are subjected to translational repression as an added layer of gene regulation, whereas others go on to increase their translation during embryogenesis. Applying *Z*-ratio analysis, we observed in total 367 canonical ORFs with significantly changed translational efficiency between the maternal and the early zygotic translatomes (Fig. 4c), of which 231 where downregulated (including 62 *Dscam* ORFs) and 136 upregulated, showing RNA expression patterns preferentially registered as "maternal" or "ubiquitous" at this stage (Fig. 4d, Additional file 1: Fig. S2C). Further, the maternal to zygotic transition in translation was also highlighted by the correlation between qualitative and quantitative changes in ribosomal binding. In total, 508 maternal-only canonical ORFs appeared translated (framed) in eggs but not in 0–8-h AEL embryos (Fig. 3b), undergoing a significant reduction in ribosomal binding efficiency during this period (as shown by TE, Fig. 3b insert). Reciprocally, 212 ORFs were just ribosome-bound in oocytes, but fully translated in early embryos, in correlation with an increase in their TE (Fig. 3b, insert). Finally, 1800 ORFs were translated in our 0–8-h AEL samples but did not yield ribosomal FPs in eggs, presumably corresponding to newly expressed, zygotic-only ORFs (Fig. 3b).

We hypothesized that a detailed comparison between S2 cells and late embryos (from where the S2 macrophage-like line was isolated [28]) might reveal the S2 cell-specific

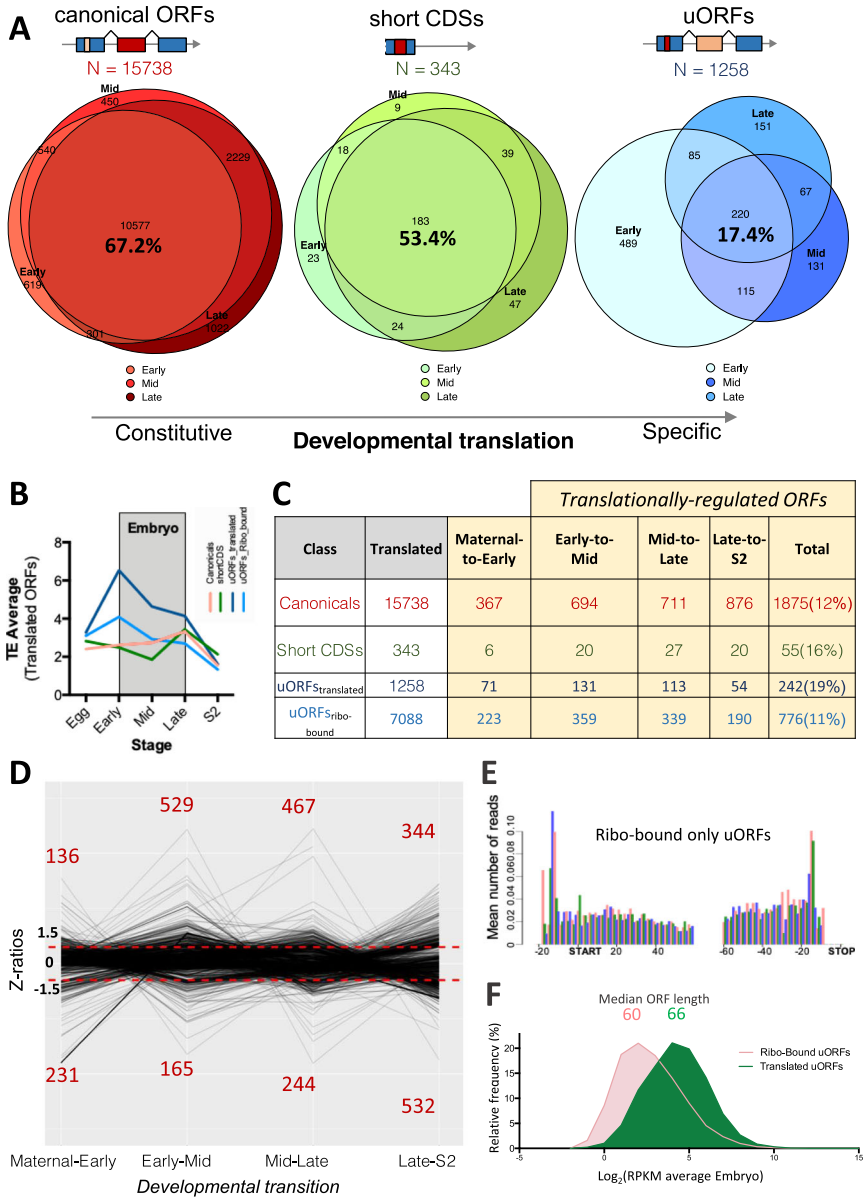

Fig. 4 Regulated translation during embryogenesis. **a** Number of transcribed, ribosome-bound, and translated ORFs per class across *Drosophila melanogaster* embryonic development, showing stage-specificity in ON/OFF translation patterns. **b** Changes in average translation efficiency, per ORF class, across analyzed developmental time-points. **c** *Z*-score ratio analysis of translation efficiencies (TE) during successive developmental stages pinpointing ORFs subjected to significant developmental modulations of translation across ORF classes. **d** Number of developmentally regulated canonical ORFs subjected to downregulation (bottom, *Z*-ratio ≤ − 1.5) or upregulation (top, *Z*-ratio ≥ 1.5), per developmental transition. **e** Metagene plot for codon framing of 32-nt ribosome footprints across reproducibly "ribosome-bound-only" uORFs, showing a lack of three-nucleotide periodicity (framing) similar to that of mRNA-Seq (compared with translated uORFs in Additional file 1: Fig. S3C). **f** Distribution of $\log_2(RPKM^{FP})$ values of ribosome-bound-only uORFs (red) and translated (framed) uORFs (green). Median nucleotide sizes of each pool are indicated above

translatome profile. Most S2 ORFs seemed translated in both ($N = 6108$), but we observed 369 ORFs translated in S2 cells but not in late embryos, whereas 8021 ORFs appeared translated in late embryos but not in S2 cells (Additional file 1: Fig. S2D). In addition, 532 embryo-translated ORFs showed significantly lowered TE in S2 cells and

344 significantly raised TE according to *Z*-ratios (Fig. 4d). In principle, these differences might reveal a macrophage-like regulatory state, an adaptation to culture, or just a sampling issue (see "Discussion"). Interestingly, we note that the S2 FPs correlate more closely with early than with late embryogenesis FPs (Fig. 3a), perhaps reflecting a "de-differentiation" of the S2 macrophage precursors in cell culture.

uORFs as translational regulators

We obtain highly reproducible Poly-Ribo-Seq signal average for uORFs across biological replicates in the embryo ($r = 0.95$, Additional file 1: Fig. S3A), indicating that most uORFs are bound by ribosomes at specific levels across embryogenesis. Indeed, a majority (72%, or 8346 of 11,710) of embryo-transcribed uORFs show reproducible ribosome binding during embryogenesis ($RPKM^{FP}>1$ in two replicates, Fig. 2f, Table S4). However, only 15% of these ribosome-binding events (1258) show reliable translation signal, as measured by the binomial framing statistic (Fig. 2f, Additional file 1: Fig. S3C), suggesting that a majority of uORFs in *Drosophila melanogaster* might not have a peptide-productive role (Fig. 4e).

Eukaryotic uORFs can act as cis-translational repressors of a canonical main ORF (mORF) located downstream in their transcripts [50]. This regulatory role has been extrapolated to the transcriptome-wide level as the main function of uORFs [51–54]. Amongst canonical protein-coding genes, we saw significant shifts in translational regulation across embryogenesis in hundreds of ORFs ($Z\text{-ratio} \geq |1.5|$) in both early-to-mid (694 ORFs) and mid-to-late embryogenesis (711 ORFs; Fig. 4c). However, a much smaller number of uORFs exhibited significant variation in translation *Z*-ratios (131 and 113 uORFs, respectively; Fig. 4c), making it unlikely that translated uORFs underpinned most translational changes in canonical ORFs. However, the cis-regulatory role of uORFs is in principle based on ribosome binding and it is mostly independent of the peptide being produced [50, 55, 56], so it could conceivably be carried out by non-productive ribosomal binding too. Therefore, we extended our analysis to include non-framed, ribosome-bound-only uORFs as well (i.e., $RPKM^{FP}>1$ but framing $p > 0.01$; Fig. 2a, f), and this provided a combined pool of 8346 ribosome-associated uORFs (Fig. 4c).

Previously, a negative correlation between the number of uORFs within a 5′ leader and the TE of the main ORF was found in zebrafish genome-wide studies [52] and across vertebrates [53]. We wondered if this was applicable to our 8346 ribosome-associated uORFs. This would indicate that this negative effect is indeed mediated by uORF ribosomal activity and not a function of other 5′ leader features [53]. However, we do not see an effect of ribosome-associated uORF number on mORF TE. Instead, we observe stable average TE across mORFs carrying different numbers of uORFs in their 5′ leaders (Fig. 5a).

Nonetheless, it would be possible that a minority of uORFs might have important cis-effects on their mORFs. We looked for this "uORF onto mORF" regulation by asking if the 1018 uORFs with significantly changed TE across stages (242 translated and 776 ribosome-bound-only, Fig. 4c) also show simultaneous significant TE changes in their cis-associated mORF. We found this not to be the case: 80% (817 of 1018) of uORFs significantly regulated do not have mORFs with significantly changed TE *Z*-

ratios (≥ |1.5|) and vice versa (Fig. 5b). Yet, 9 uORFs correlated negatively with their canonical ORF (i.e., as the uORF TE went up, the mORF TE went down, Fig. 5b), suggesting a negative cis-translational regulatory role for these uORFs. Another 19 uORFs displayed reduced TE while their respective mORF TEs were upregulated, which is compatible with such a negative regulatory role. However, 134 uORFs showed coordinated changes with their mORF (i.e., uORF and mORF went both up (86) or both down (48) significantly (Fig. 5b), and this positive correlation extended to the TEs of all 8346 ribosome-associated uORFs and their mORFs ($r = 0.47$ Fig. S3B). This positive correlation could indicate that uORFs act as either (a) positive translational regulators, (b) constant "brakes" or negative regulators that reduce but do not overcome mORF translation, or finally, (c) passive "bystanders" that are ribosomal-bound by virtue of being present in the 5′ leader of a transcript containing a highly translated mORF. In addition, a combination of these roles is also possible. Interestingly, the total range of TE variation diminishes as the number of ribosome-bound uORFs increases within a given mORF 5′ leader (Fig. 5a), with canonical ORFs that contain no uORFs in their 5′ leaders displaying twice the coefficient of variation as those with one or more ribosome-bound uORFs (Fig. 5a, inset). This suggests a cumulative role for uORFs in the maintenance of mORF translational efficiencies, possibly acting as buffers of variation in mORF translation. A buffering effect could also explain the higher TE averages amongst uORFs than amongst canonical ORFs (Fig. 4b), even though the proportion of ORFs with significantly changed TE was identical (1018 of 8504 uORFs: 12% vs. 12% canonicals, Fig. 4c, but see also below).

We observed that 45% of canonical ORF 5′ leaders contain uORFs, a number in line with observations across metazoans [25, 29, 57]. Interestingly, this number is significantly lower in short CDSs (25%). This could be explained by the difference in 5′ leader average lengths between these two classes (306 nt versus 194 nt). Indeed, the number of uORFs is positively correlated with the length of the 5′ leaders across all annotated mRNAs ($r = 0.668$, Fig. 5c), indicating that uORFs tend to accumulate more often in longer 5′ leaders. The positive correlation between uORF number and 5′ leader length supports the idea that uORFs appear randomly in mRNAs [29].

Regardless of whether uORFs act as cis-translational regulators or not, an important number (1258, Fig. 2f) show framing, i.e., appear to be translated into peptides. On average, $RPKM^{FP}$ (Fig. 4f) and TE (Fig. 4b) was higher in these framed uORFs and changed more dramatically across stages than on ribosome-bound-only uORFs (note steeper slope in Fig. 4b and higher total percentage of significant translational regulation in Fig. 4c). The regulation of translated uORF expression was not only quantitative, but included qualitative changes as well (i.e., from ribo-bound-only to framed, correlating with an increase in 32-nt FPs mapping to the uORF), and transcriptional regulation, altogether producing that 83% of uORFs were translated in a stage-specific manner (Fig. 4a). Thus, regulated expression seemed a feature of uORF translation.

In the absence of proteomic data to corroborate whether framed uORFs do indeed produce peptides, we checked other bioinformatics markers of coding potential. We observed that translated uORF start codon contexts show significantly better Kozak scores [50] than ribosome-bound-only uORFs (Fig. 5d). Further, the sequence conservation of translated uORFs is intermediate between transcribed uORFs and canonical ORFs (Fig. 5e). The amino acid usage produced a similar intermediate picture (Fig. 5f).

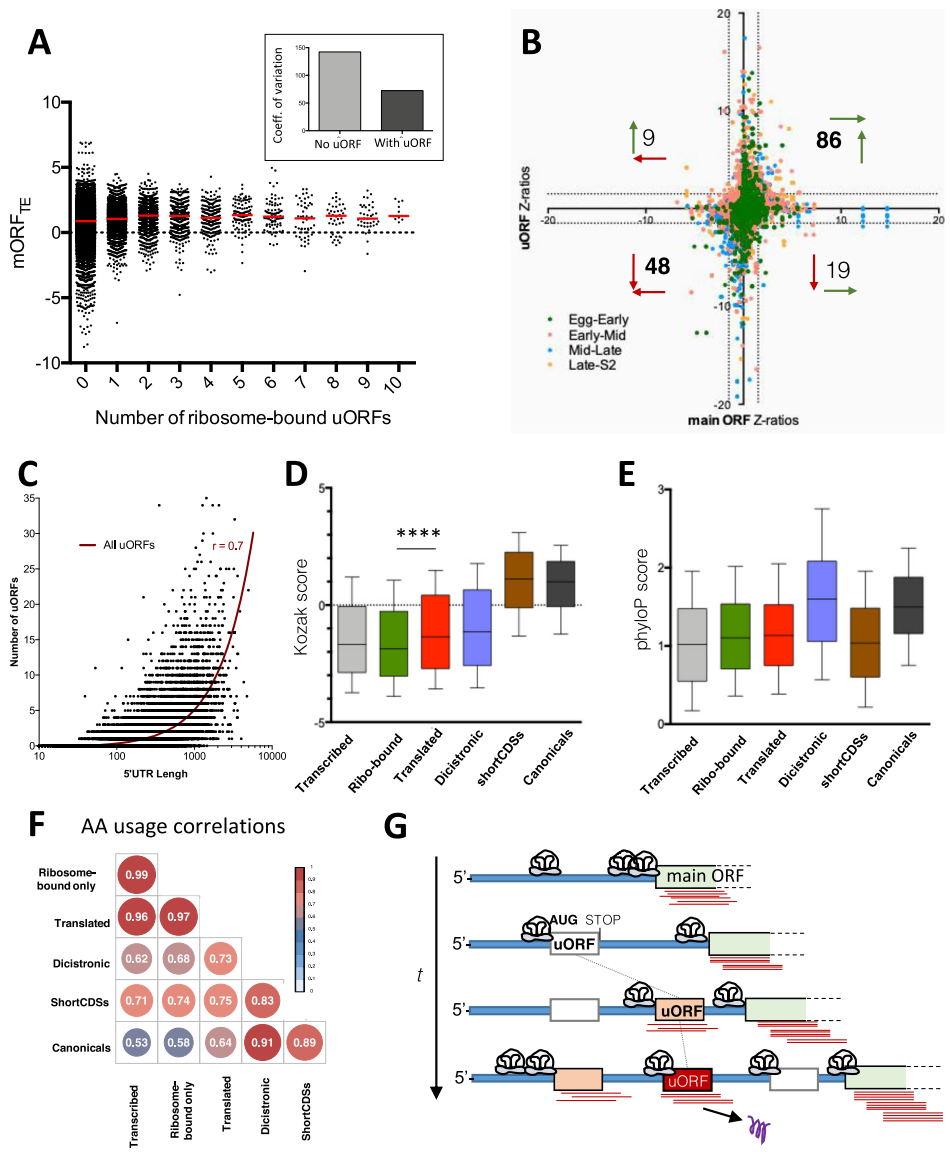

Fig. 5 uORFs as translational regulators. **a** Variation in canonical TE levels as a function of the number of cistronic uORFs showing ribosomal binding. **b** Comparison of translational changes between uORFs and their canonical mORFs across developmental stages. Numbers denote the number of *Z*-ratio significant uORF-mORF pairs in each quadrant. Vertical arrows: uORF translational regulation. Horizontal arrows: canonical mORF translational regulation. Green arrows: upregulated; red arrows: downregulated. **c** Correlation between the number of uORFs per mRNA and 5′UTR length (Spearman's $r = 0.7$). **d** Comparison of Kozak-context scores between distinct ORF classes and uORF subclasses. Graphs display 10–90 percentile range. Mann-Whitney $p < 0.001$. **e** Comparison of phyloP conservation scores (27-way alignments) between distinct ORF classes and uORF subclasses. Graphs display 10–90 percentile range. **f** Correlations in amino acid usage across distinct uORF subclasses, short CDSs, and canonical ORFs; values denote pairwise Spearman's *r*. **g** Model for uORF function and evolution. 5′ leaders in blue. t: evolutionary time. Red lines denote ribosomal binding signal; white boxes correspond to novel uORFs; salmon boxes correspond to ribosomal-bound uORFs; red boxes correspond to translated uORFs. Purple scribble denotes produced peptide

These observations might suggest that translated uORFs either produce a specific type of peptide or that translated uORFs are in an evolutionary transition into coding-ness [29]. One end result of such a transition would be to produce a di-cistronic coding

gene (uORF plus mORF), and it is interesting that genes annotated as such by FlyBase display similar conservation to canonical ORFs (Fig. 5e), but amino acid usage intermediates between these and translated uORFs (Fig. 5f).

Discussion

Translation vs. peptide function

Protein translation is the most expensive gene activity undertaken by a cell, one order of magnitude more costly than either replication or transcription, and costly enough to have a selective impact [58]. However, it is important to disentangle the act of productive translation from its biological purpose. It is usually expected that translated ORFs must produce a peptide or a protein with canonical function. However, a translated peptide could have other biological activities, [59] or no peptide function at all, such as the act of translation itself could be the biologically relevant function of the encoding ORF (as in regulatory uORFs, or in canonical ORFs during nonsense-mediated decay) [55, 60]. Ribo-Seq, and other techniques such as proteomics, do not reveal protein and peptide function, only its expression. Protein function is suggested by other studies (such as sequence conservation) and ultimately tested experimentally.

smORF translation is difficult to ascertain, but has been repeatedly detected to occur at a large scale by Ribo-Seq, including in this study. Different types of studies and analyses yield numbers always in the thousands per genome [45]. These numbers are a small fraction of the smORFs in the genome (about 5%), yet it is large enough (about a 10% of extra coding genes to add to canonical genes) to present a challenge to our understanding of genomes and their function. What can be the biological purpose of this 10% "crypto-translatome"? Due to their shortness, it is challenging to determine smORF function from their sequence. GO analysis of the longer, and annotated, sCDS smORF class, and analyses of their aminoacid composition [23, 29], suggested roles related to cell membranes and organelles. However, shorter smORFs such as uORFs and those found in lncRNAs (called lncORFs) are not annotated and do not display known protein domains, nor homologies to characterized proteins, that could suggest a peptide function. Their AA composition is intermediate, yet distinct from canonical proteins and random sequences [29]. Nonetheless, important functions for lncORF and uORF peptides have been proven experimentally in a few cases [30, 39, 61–64], as well as a cis-regulatory, "non-coding" function for uORFs [50, 55, 56].

Ribosomal binding versus productive translation

Our results suggest a distinction between ribosomal-binding-only and productive translation. We observed that about 60% of transcribed uORFs show abundant, yet not framed, ribosomal binding. This is unlike canonical ORFs and short CDSs, where only 6% and 20% respectively show ribosomal association not corresponding to productive translation. This raises the possibility that either much of ribosomal binding in uORFs serve purposes other than peptide production, or that it reflects "translational noise." In view of the energetic and selective costs involved, it is unlikely for the cells to allow for their resources to just be "lost in translation." An explanation might be suggested by the maternal translatome.

Maternal translation

The maternal to zygotic transition corroborates that "ribosomal-binding-only" without framing is a different state than fully fledged translation. Our data revealed ORFs "poised" for translation between unfertilized eggs and early embryos. A total of 212 canonical ORFs switch from unproductive ribosome-bound-only to productive translation, while increasing their translational efficiency (Fig. 3b). In other words, the quality of their ribosomal association changes from low affinity (low TE) and not following a particular frame, to higher-affinity, frame-linked productive translation. To produce non-framing, ribosomal binding must happen in overlapping frames, "shift" between frames, or include "scanning" 40S subunits and/or proofreading ribosomal units involved in NMD and not productively reading codons. Interestingly, we observe one-frame shift of ribosomes at STOP codons (from blue frame "2" to red "0", Fig. 2b, e), also noticeable in other Ribo-Seq studies [21], which may reflect a conformational change or the ribosome pausing for a time at its A site while disengaging from the mRNA. Under weak translation conditions (indicated by lower $RPKM^{FP}$, Fig. 4f), in some ORFs elongation may proceed more slowly (perhaps due to a failure to acquire AA-loaded tRNAs or elongation factors). It has been shown that prolonged ribosomal stalling plus queuing can produce both premature termination [65], and use of non-AUG START codons in a different frame [66]. Either effect could produce a frame-shifted Ribo-Seq signal towards frames "0" and "1" as we observe, distorting the main ORF signal in frame "2" (Fig. 4e). In other words, ribosome-bound-only ORFs could be failing to produce enough START to STOP translation, while increasing a frameshifted, noisy ribosomal binding. This effect could be temporary (as in the maternal to zygotic transition) or a constitutive feature of the ORF (as in most uORFs, see below).

Function of uORFs

In our previous study [23], we estimated that 34% of *Drosophila* uORFs were translated in S2 cells, on the basis of a high $RPKM^{FP}$. Interestingly, our improved methods reveal two uORF pools, whose $RPKM^{FP}$ values overlap yet having different averages (Fig. 4f), and different TE (Fig 4b) and framing (Fig. 4e and Fig. S3C). The percentage of ribosome-bound uORFs in *Drosophila* embryos (71%) is similar to recent estimates [54], but the percentage of actually translated uORFs in our samples is only 11%. The significance of non-productive ribosomal-binding in 7088 (60%) of transcribed uORFs is intriguing and is not explained by the 302 (4.2%) overlapping translated uORFs in another frame. The genome-wide lack of negative effect of ribosome-bound uORFs on mORF TE, the mild positive correlation between such uORFs and mORFs, and the correlation between 5′ leader length and number of uORFs all seem to suggest that most uORFs are a random and function-less by-product of 5′ leader sequence variation, also suggested by the random-like size distribution of uORFs [29]. Yet, this does not exclude that some uORFs could have a cis-regulatory or a peptide-producing role. Different mechanisms have been proposed to mediate a cis-regulatory function, including ribosome stalling and disengagement at uORFs, a reduction in translational re-initiation on the downstream mORF AUG codons, and the triggering of mRNA decay [50, 56, 67], all leading to a decrease in mORF translational efficiency. Positive regulatory roles are also possible, with uORFs recruiting ribosomes and thus increasing the likelihood that

they reach mORFs via reinitiation, and/or imposing an appropriate ribosomal spacing (reducing ribosome collisions and thus improving mORF TE and fidelity [65, 66, 68]). We propose a combined role, with most uORFs acting as "translation buffers" that would recruit ribosomes (a relatively scarce yet valuable cellular resource [68, 69]) and pass them on to their mORF at a constant rate, while discarding ribosomes in excess (Fig. 5g). A passive, low-level uORF buffer effect at the genome-wide level is consistent with less variation of mORF TE correlating with more ribosome-bound uORFs (Fig. 5a); and also with studies in yeasts [70, 71], vertebrates [25], and humans [72], where a genome-wide positive correlation between the TE of individual uORFs and their cognate mORFs is also observed. A function of uORFs as buffers could also generate stalling, queuing, termination, and frameshifted starts [66], and explain their frameshifted FP signal.

A mildly beneficial buffering role could explain the ubiquity of uORFs in otherwise canonical mRNAs. The minority of uORFs showing peptide-productive translation could be a tolerable by-product of this buffering role, which would otherwise be mostly mediated by the much more prevalent uORFs with non-productive ribosomal-binding-only. Interestingly, we observe subtly different Kozak scores, conservation levels, and amino acid usage for translated uORFs, which seem to fit into a stepped continuum ranging from transcribed-only uORFs to canonical ORFs. It is tempting to speculate that, just as ribosomal-bound-only in canonical ORFs may be a developmental transitory state during the maternal to zygotic translation, ribosome-bound-only uORFs may be an evolutionary transitory state poised for a transition to full coding function. In this way, the 5′ leaders of canonical genes would act as "proto-gene nurseries" offering a tolerant environment for random ORF creation (Fig. 5g), while providing coding mRNA features such as poly-A tails, splicing-related and translationally related RNA processing and stability, allocation to polysomes, and a specific transcriptional pattern. Dicistronic genes could be another step in this transition from inert or regulatory uORF to coding ORF (although dicistronics can also emerge from the fusion of two pre-existing canonical ORFs). Interestingly, short CDSs, which in most aspects studied here (RPKM reproducibility, translation percentage, MassSpec detection, embryonic stage specificity, prevalence of translation regulation, amino-acid usage) are intermediate between translated uORFs and canonical ORFs, do not show intermediate conservation (Fig. 5e). The lower sCDS conservation may suggest that sCDSs are also involved in another evolutionary process, perhaps another and more dynamic stepped continuum leading to coding-ness via lncRNAs [29, 73].

Further developments—single cell translatomics

Despite the improvements to our Poly-Ribo-Seq protocol, and the large number of reads that we have collected, it is possible that not all embryonic translation has been detected in our samples. Lowly abundant or transient mRNAs may not be adequately detected in our 8-h developmental windows, or may be drowned out by noise and not be able to achieve RPKM > 1. This could be especially true of mRNAs expressed in a few cells of the embryo, a situation that must arise often towards the end of embryogenesis, and could be prevalent in differentiated tissues. Our macrophage-like S2 cell line data allows for an approximation to this issue, and the problem of integrating and

comparing Ribo-Seq and genomic data from whole embryos and cell lines. Direct comparison between late embryos and S2 cells creates a sampling issue. The embryonic counterparts of S2 cells could be only a few cells amongst many embryonic cell types and some 30,000 cells overall [74, 75]. Thus, the FPs from S2-related ORFs might not be adequately detected in whole embryo extractions (thus yielding a low or no RPKM), whereas the same FPs will be well-detected in the monotypic S2 cell culture. For example, we showed the specific expression (by in situ hybridization and RT-PCR) and function (by observation of mutant phenotypes in embryonic macrophages) of the *hemotin* smORF gene in the *Drosophila* embryo [76]. Yet our RNA-Seq and Ribo-Seq data, and modENCODE RNA-Seq would indicate that *hemotin* is neither transcribed nor translated during embryogenesis, while the same data show vigorous expression in S2 cells ($RPKM^{FP} = 58.6$ and framing $p = 6.44 \times e10^{-4}$) and hence a TE *Z*-ratio significantly increased. This sampling issue can contrive an artefactual increase of expression and translation (RPKM and TE) of *hemotin* and similar cell-specific transcripts in cell lines [77] when compared with whole embryos or organs. However, a lowering of RPKM or TE in S2 cell cultures in 532 ORFs (Fig. 4d) cannot be due to this sampling issue and must reflect specific translational repression in a S2 "cell fate," following either a macrophage-like fate, or a cell-in-culture fate. Thus, Ribo-Seq data from cell lines can reveal genes whose translation declines during specific cell differentiation programs. Ideally, it would be best to take RiboSeq to this single cell level. The demand for high levels of starting material (in the region of μg) is the largest barrier to single-cell Ribo-Seq, but it could be overcome by the combination of ribosomal immunoprecipitation (TRAP) (which requires even more starting material, in the region of mg [78, 79]), with further improvements to the core polysome purification protocol [80].

Conclusions

The likelihood of translation can be determined in Ribo-Seq data by comparing the frequency of reads in frame against a simple binomial probability. Using this metric and a standard RPKM>1 filter for ribosomal binding, almost 16,000 canonical ORFs, more than 300 sCDSs, and more than 1200 uORFs appear translated during the embryonic development of *Drosophila melanogaster*. Translation levels appear more reproducible than transcription, yet highly correlated with it, together yielding highly constant expression of canonical ORFs and sCDSs, and higher temporal specificity of uORFs. However, 11–19% ORFs show specific regulation of their translation, including "poising" at the maternal to zygotic transition. Finally, a large pool of near 7000 uORFs also show ribosomal binding without achieving either productive translation or a significant regulatory input on downstream ORFs. These and other data suggest that, in general, 5′ leaders with uORFs act as "translation buffers" at the gene level (helping downstream ORFs to stabilize translation levels) and as "proto-gene nurseries" at the genomic level (providing a favorable environment for random ORF creation).

Materials and methods

Poly-Ribo-Seq and RNA-seq procedures

The RNA-Seq and Poly-Ribo-Seq experiments were conducted as previously described in Aspden et al. [23] with the following modifications. Embryos were flash frozen,

turned into powder using a pre-chilled pestle and mortar and homogenized in Lysis Buffer at 4 °C for 20 min. Pre-clarified lysates were processed for mRNA isolation and fragmentation and polysome separation and digestion as previously described. Digested polysome samples were concentrated by centrifugation and subsequently loaded into a 1M Sucrose Cushion and centrifuged at 70,000*g* to pellet the monosomes. The 50 nt fragmented mRNAs and 28–34-nt footprints were isolated from 10% denaturing acrylamide gels and subsequently T4 PNK treated for library preparation.

The NEBNext Small RNA Library Prep Set for Illumina (NEB, E7300) was used following the manufacturers' instructions with small modifications as follows. After the 3′ adapter ligation step, ribosome footprints were rRNA depleted using biotinylated DNA fragments and oligos from rRNAs and going through two consecutive rounds of subtractive hybridization (except for the T 0-8 h sample having only one round) using MyOne Streptavidin C1 Dynabeads (Invitrogen). After ethanol precipitation, samples were processed as NEB guidelines. Libraries were isolated from 10% non-denaturing gels by size selection and their quality and quantity measure by DNA high sensitivity chips (Agilent) with Bioanalyzer and Qubit.

Footprint sequence alignment

Ribosomal footprints were filtered (PHRED quality ≥ 33) and clipped for adapters using the FASTX-Toolkit. The resulting reads were aligned to a FASTA file containing rRNA, tRNA, snoRNA, and snRNA from Flybase Release 6.13 annotations using Bowtie, discarding the successful alignments and collecting all unaligned reads. Unaligned reads were then mapped to the FlyBase *D. melanogaster* reference genome (Release 6.13) using HISAT2 with default options. Due to the very low frequency of multimapping reads, we retained all genome-aligned reads for further analyses.

ORF selection and identification

All annotated coding sequences (CDSs) in FlyBase, r6.13 were retrieved and divided into groups of either longer than or shorter than 303 nucleotides (100 AA). CDSs longer than the cut off were assigned to the canonical ORF category. CDSs shorter or equal to 303 nt but arising from the same locus as canonicals (short-isoform smORFs, [29]) were discarded from our analysis. The remaining CDSs of less than 100 AA which arise from independent genomic loci, were assigned to the short CDS category [29]. We identified uORFs longer than 10 AA with an AUG start codon followed by an in-frame stop codon within the annotated 5′ leaders of all transcripts annotated as protein-coding in FlyBase r6.13, using the emboss *getorf* program. Next, all ORFs with fully redundant CDS coordinates were identified, and duplicates were discarded within each class (canonicals, short CDSs, and uORFs). Additionally, we discarded uORFs overlapping with any annotated CDS, or fully included within a longer, in-frame uORF, but kept uORFs overlapping in different frames. uORFs in dicistronic transcripts were then added to the uORF set, based on a full overlap with annotated ORFs contained within the 5′ leader of an immediately downstream annotated CDS.

Ribosomal footprint analysis

Relative ribosome density ($RPKM^{FP}$) was measured by counting Ribo-Seq reads of 26 to 36 nt overlapping each ORF using the R package Rsamtols (absolute ribosome density), and scaling this number by ORF length and the total number of genome-aligned reads [2]. Total mRNA expression ($RPKM^{RNA}$) was ascertained using the same ORF coordinates and applying the same method on RNA-Seq measurements. Correlation analyses across replicates were measured by Spearman's rho on ORF RPKM values.

Detection of framing and individual translation events

To analyze framing, ribosome footprints (FPs) were first aligned to an artificially constructed transcriptome set, consisting of all ORFs analyzed in this study as well as their surrounding regions − 18 nucleotides upstream and + 15 nucleotides downstream, thus allowing for the alignment of full ribosomal reads spanning the START and STOP codons. The resulting alignments were analyzed using the R package RiboSeqR to extract the global framing patterns across ribosome footprint lengths for three pools of RPFs: all pooled embryonic RiboSeq samples, S2 cell samples, and pool of mature oocyte and activated eggs RPFs (Kronja et al. [18], cycloheximide^{+}). The main read length and its most overrepresented frame in each sample was then used to evaluate the framing of each ORF in each Ribo-Seq sample. For each ORF, we then compared successful framed reads (those matching global framing patterns) with total mapped reads using a binomial test, implemented by the cumulative *pbinom()* function in R to measure the probability of obtaining the observed framing (or better), when compared to the expectation by random (1 in 3 probability). This was implemented using the expression *pbinom([framed_reads]-1, [all_reads], 1/3, lower.tail = FALSE).* For binomial tests of ORF translation, different biological replicates (if available) were merged into a single one. Although the precision of the binomial probability method improves with the amount of evidence (i.e., higher number of reads can achieve very low p values, or very high probabilities of being translated, Fig. 2c), it can still achieve results with minimal information, which is especially relevant for smORFs as their short size accrues less reads in sequencing experiments when compared to canonical ORFs with the same RPKM range (Figs. 1 and 2e). Thus, a minimum of 5 reads is required to obtain $p < 0.01$ in a given stage if all reads are in frame, or 7 or more reads if any is out of frame [38].

Comparison with other Ribo-Seq and RNA-Seq datasets

We obtained previously published [18] fastq files containing Cycloheximide^{+} Ribo-Seq reads from mature oocytes and activated eggs from the NCBI Gene Expression Omnibus (accessions SRR1039770 and SRR1039771) as well as the corresponding RNA-Seq experiments (SRR1039764 and SRR1039765), all deposited under accession number GSE52799. The activated egg and oocyte reads were merged to obtain Ribo-Seq and RNA-Seq "unfertilized egg" samples, which were analyzed employing the same pipeline as our own data. In the case of S2-cell data, we merged our previously published Ribo-Seq (accessions SRR1548656, SRR1548657, SRR1548658, SRR1548659) and RNA-Seq datasets (SRR1548660-SRR1548661) [23] and added a novel round of Poly-Ribo-Seq for a total of 382,232,123 raw reads and 23,710,309 genome-aligned footprints. We obtained modENCODE RNA-Seq 2-hourly

embryonic-staged datasets from FlyBase and consolidated them into our 8-h-long stages (0 h–8 h, 8 h–16 h, and 16 h–24 h AEL).

Comparison between RNA-seq and mass spectrometry datasets

We compared the mass spectrometry-based developmental proteome of *Drosophila melanogaster* from [26] with our own sequencing data by first averaging the imputed log2 LFQ intensity per protein-group across samples spanning embryonic windows matching those of our timed collections (0 h–8 h, 8 h–16 h, and 16 h–24 h AEL). Finally, we normalized all datasets and measured the correlation between the average protein intensities for all identified proteins and our own gene expression measurements using Spearman's rho.

Translation efficiency and translational regulation analysis

Translation efficiency (TE) per ORF was calculated as the fraction of ribosome-density (Poly-Ribo-Seq RPKM) over the total mRNA read density (RNA-Seq RPKM). To detect significant events of translational regulation, we performed *Z*-ratio calculations of TE variation as per Cheadle et al. [48]. First, TE values were normalized by calculating the *Z*-score of each ORF in each sample:

$$Z\text{score}_{\text{ORF}} = \left(\text{TE}_{\text{ORF}} - \text{TE}_{\text{sampleMean}}\right)/\text{TE}_{\text{sampleSD}}\right)$$

Second, *Z*-ratios across contiguous developmental stages, e.g., early and mid embryogenesis, were calculated for each ORF:

$$Z\text{–ratio} = \left(\text{AverageZScore}_{\text{Mid}} - \text{Average } Z\text{score}_{\text{Early}}\right)/\left(Z\text{score}_{\text{Mid}} - Z\text{score}_{\text{Early}}\right)_{\text{SD}}$$

Third, the cut-off of ≥ |1.5| was used to identify biologically significant differences [47, 48] between the two contiguous stages.

Gene expression enrichment analysis

We used the unique gene identifiers (FBgn) for genes identified as translationally regulated across stages (see "Translation efficiency and translational regulation analysis") to calculate the embryonic tissue-expression enrichment using the Intermine tool (Flymine), which uses ImaGO terms associated with expression patterns deposited in the BDGP database. The enrichment for each gene set was measured against a background of all *Drosophila melanogaster* genes with expression information and expressed as a *p* value for each significantly enriched tissue, after the Benjamini-Hochberg multiple-comparisons correction.

Conservation analysis

The pre-computed phyloP nucleotide scores for 27-way insect multiple sequence alignments were downloaded from UCSC as bigWig files. We then used the UCSC bigWigAverageOverBed package to compare phyloP scores with BED files containing genome-coordinates for each ORF in our set, obtaining average phyloP scores per ORF.

Amino acid composition analysis

We used a previously published own script (*aa_composition.pl*, [23]) to calculate the observed or predicted amino acid compositions of our different sets of ORFs.

Kozak-context scoring per ORF

For all canonical ORFs, we extracted the nucleotide composition around—but excluding—the annotated START codon (− 5 to 6, excluding positions 1–3). For each position, we then calculated the log odds ratio between observed and background nucleotide frequencies. This provided a scoring table of position-specific nucleotide frequencies from *bona fide* Kozak contexts with which to score individual ORFs in all classes. The final score per ORF was then obtained by adding the individual position-specific values for all observed nucleotides.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-02011-5.

Additional file 1. Patraquim_Supplemental_Information_Additional File 1.pdf: Supplementary Figures and Tables.

Additional file 2. Patraquim_canonical_ORFs_Additional_File_2.csv: transcriptional, ribosome-binding, framing and translational regulation analysis results for canonical ORFs.

Additional file 3. Patraquim_shortCDSs_Additional_File_3.csv: transcriptional, ribosome-binding, framing and translational regulation analysis results for shortCDSs.

Additional file 4. Patraquim_uORFs_Additional_File_4.csv: transcriptional, ribosome-binding, framing and translational regulation analysis results for uORFs.

Additional file 5. Review history.

Acknowledgements

We thank Rose Phillips for excellent technical assistance, Unum Amin for transfection and confocal microscopy experiments, and Emile Magny, Sarah Bishop, and Sarah Newbury for discussions and comments to the manuscript. We also thank Terry Orr-Weaver and Stephen Eichhorn for providing information on the Kronja et al. data.

Peer review information

Tim Sands was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 5.

Funding

This work was funded by University of Leeds start-up funds to JA and by Grants from the BBSRC (Ref BB/N001753/1), and the Spanish MINECO (Ref. BFU2016-077793P) to JPC. These bodies had no role in the design of the study or the collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

All data generated in this study have been deposited to the GEO depository (GSE147619) [81].

Authors' contributions

PP contributed to the conception and overall design of the work, bioinformatics analysis and interpretation of Poly-Ribo-Seq, proteomics and genomics data; writing of software; and drafting of the manuscript. MASM contributed to the design of Poly-Ribo-Seq experiments and acquisition of Poly-Ribo-Seq data. JIP contributed to the acquisition, bioinformatics analysis, and interpretation of Poly-Ribo-Seq data; design and acquisition of tagging data; and substantial revision of the manuscript. JLA contributed to the interpretation of Poly-Ribo-Seq data, design and acquisition of tagging data, and substantial revision of the manuscript. JPC contributed to the conception and overall design of the work, interpretation of tagging and Poly-Ribo-Seq data, and drafted the manuscript. The authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Present Address: Centro Andaluz de Biologia del Desarrollo, CSIC-UPO, Seville, Spain. [2]Brighton and Sussex Medical School, Brighton, East Sussex, UK. [3]School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK.

Received: 27 September 2019 Accepted: 8 April 2020
Published online: 29 May 2020

References

1. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011;147:789–802.
2. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324:218–23.
3. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, Wills MR, Weissman JS. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. 2014;8:1365–79.
4. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, Landthaler M, Obermayer B, Ohler U. Detecting actively translated open reading frames in ribosome profiling data. Nat Methods. 2016;13:165–70.
5. Mumtaz MA, Couso JP. Ribosomal profiling adds new coding sequences to the proteome. Biochem Soc Trans. 2015;43:1271–6.
6. Jang C, Lahens NF, Hogenesch JB, Sehgal A. Ribosome profiling reveals an important role for translational control in circadian gene expression. Genome Res. 2015;25:1836–47.
7. Spriggs KA, Bushell M, Willis AE. Translational regulation of gene expression during conditions of cell stress. Mol Cell. 2010;40:228–37.
8. Sendoel A, Dunn JG, Rodriguez EH, Naik S, Gomez NC, Hurwitz B, Levorse J, Dill BD, Schramek D, Molina H, et al. Translation from unconventional 5′ start sites drives tumour initiation. Nature. 2017;541:494–9.
9. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. Science. 2016;352:1413–6.
10. Stadler M, Fire A. Conserved translatome remodeling in nematode species executing a shared developmental transition. PLoS Genet. 2013;9:e1003739.
11. Gilbert SF: Developmental biology. Fifth edition. Sunderland: Sinauer Associates; 1997; p. 1–918.
12. Zalokar M. Autoradiographic study of protein and RNA formation during early development of Drosophila eggs. Dev Biol. 1976;49:425–37.
13. Qin X, Ahn S, Speed TP, Rubin GM. Global analyses of mRNA translational control during early Drosophila embryogenesis. Genome Biol. 2007;8:R63.
14. Yoshikane N, Nakamura N, Ueda R, Ueno N, Yamanaka S, Nakamura M. Drosophila NAT1, a homolog of the vertebrate translational regulator NAT1/DAP5/p97, is required for embryonic germband extension and metamorphosis. Develop Growth Differ. 2007;49:623–34.
15. Bate M, Martinez-Arias, A. (Ed.). The development of *Drosophila melanogaster*. New York: Cold Spring Harbour Laboratory Press; 1993;1:1–746.
16. Ray RP, Arora K, Nusslein-Volhard C, Gelbart WM. The control of cell fate along the dorsal-ventral axis of the Drosophila embryo. Development. 1991;113:35–54.
17. Stjohnston D, Nusslein-Volhard C. The origin of pattern and polarity in the Drosophila embryo. Cell. 1992;68:201–19.
18. Kronja I, Yuan B, Eichhorn SW, Dzeyk K, Krijgsveld J, Bartel DP, Orr-Weaver TL. Widespread changes in the posttranscriptional landscape at the Drosophila oocyte-to-embryo transition. Cell Rep. 2014;7:1495–508.
19. Tadros W, Lipshitz HD. The maternal-to-zygotic transition: a play in two acts. Development. 2009;136:3033–42.
20. Yartseva V, Giraldez AJ. The maternal-to-zygotic transition during vertebrate development: a model for reprogramming. Curr Top Dev Biol. 2015;113:191–232.
21. Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, Giraldez AJ: Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation**.** In EMBO J, vol. 33, 2014/04/08 edition. pp. 981–993; 2014:981–993.
22. Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, Mastrobuoni G, Rajewsky N, Kempa S, Selbach M, Obermayer B. Extensive identification and analysis of conserved small ORFs in animals. Genome Biol. 2015;16:179.
23. Aspden JL, Eyre-Walker YC, Philips RJ, Brocard M, Amin U, Couso JP. Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. eLife. 2014;3:e03528.
24. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell. 2013;154:240–51.
25. Chew GL, Pauli A, Rinn JL, Regev A, Schier AF, Valen E. Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. Development. 2013;140:2828–34.
26. Casas-Vila N, Bluhm A, Sayols S, Dinges N, Dejung M, Altenhein T, Kappei D, Altenhein B, Roignant JY, Butter F. The developmental proteome of Drosophila melanogaster. Genome Res. 2017;27:1273–85.
27. Heyer Erin E, Moore Melissa J. Redefining the translational status of 80S monosomes. Cell. 2016;164:757–69.
28. Schneider I. Cell lines derived from late embryonic stages of Drosophila melanogaster. J Embryol Exp Morphol. 1972;27:353–65.
29. Couso JP, Patraquim P. Classification and function of small open reading frames. Nat Rev Mol Cell Biol. 2017;18:575–89.
30. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol. 2007;5:1052–62.
31. Martinez-Arias A: Development and patterning of the larval epidermis of Drosophila. In The development of Drosophila melanogaster. Edited by Bate CM, Martinez Arias A. New York: Cold Spring Harbour Laboratory Press; 1993. p. 517–608.
32. Skaer H: The Alimentary Canal. In The development of Drosophila melanogaster. Edited by Bate. M, Martinez Arias.A. New York: Cold Spring Harbor Laboratory Press; 1993. p. 941–1012.
33. Miguel-Aliaga I, Jasper H, Lemaitre B. Anatomy and physiology of the digestive tract of Drosophila melanogaster. Genetics. 2018;210:357–96.
34. Bate M. The embryonic development of larval muscles in Drosophila. Development. 1990;110:791–804.
35. Landgraf M, Jeffrey V, Fujioka M, Jaynes JB, Bate M. Embryonic origins of a motor system: motor dendrites form a myotopic map in Drosophila. PLoS Biol. 2003;1:e41.

36. Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. eLife. 2013;2:e01179.
37. Calviello L, Ohler U. Beyond read-counts: Ribo-seq data analysis to understand the functions of the transcriptome. Trends Genet. 2017;33:728–44.
38. Zar JH: Biostatistical analysis. 2019.
39. Magny EG, Pueyo JI, Pearl FM, Cespedes MA, Niven JE, Bishop SA, Couso JP. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science. 2013;341:1116–20.
40. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. The developmental transcriptome of Drosophila melanogaster. Nature. 2011;471:473–9.
41. Loose CR, Langer RS, Stephanopoulos GN. Optimization of protein fusion partner length for maximizing in vitro translation of peptides. Biotechnol Prog. 2007;23:444–51.
42. Jenssen H, Aspmo SI. Serum stability of peptides. Methods Mol Biol. 2008;494:177–86.
43. Amin U. Cellular characterisation of small open reading frame function in *Drosophila melanogaster*. Doctoral Thesis. Brighton: University of Sussex, Biology; 2016.
44. Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, Potthast F, Deutsch EW, Panse C, de Lichtenberg U, Rinner O, et al. A high-quality catalog of the Drosophila melanogaster proteome. Nat Biotechnol. 2007;25:576–83.
45. Pueyo JI, Magny EG, Couso JP. New peptides under the s (ORF) ace of the genome. Trends Biochem Sci. 2016;41:665–78.
46. Zhong Y, Karaletsos T, Drewe P, Sreedharan VT, Kuo D, Singh K, Wendel HG, Ratsch G. RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints. Bioinformatics. 2017;33:139–41.
47. Quackenbush J. Microarray data normalization and transformation. Nat Genet. 2002;32(Suppl):496–501.
48. Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of microarray data using Z score transformation. J Mol Diagn. 2003;5:73–81.
49. Thomsen S, Anders S, Janga S, Huber W, Alonso C. Genome-wide analysis of mRNA decay patterns during early Drosophila development. Genome Biol. 2010;11:1–27.
50. Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene. 2005;361:13–37.
51. Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A. 2009;106:7507–12.
52. Johnstone TG, Bazzini AA, Giraldez AJ. Upstream ORFs are prevalent translational repressors in vertebrates. EMBO J. 2016; 35:706–23.
53. Chew GL, Pauli A, Schier AF. Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. Nat Commun. 2016;7:11663.
54. Zhang H, Dou S, He F, Luo J, Wei L, Lu J. Genome-wide maps of ribosomal occupancy provide insights into adaptive evolution and regulatory roles of uORFs during Drosophila development. PLoS Biol. 2018;16:e2003903.
55. Somers J, Poyry T, Willis AE. A perspective on mammalian upstream open reading frame function. Int J Biochem Cell Biol. 2013;45:1690–700.
56. Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet. 2014;15:193–204.
57. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, Schier AF. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. 2012;22:577–91.
58. Lynch M, Marinov GK. The bioenergetic costs of a gene. Proc Natl Acad Sci. 2015;112:15690.
59. Starck SR, Tsai JC, Chen K, Shodiya M, Wang L, Yahiro K, Martins-Green M, Shastri N, Walter P. Translation from the 5′ untranslated region shapes the integrated stress response. Science. 2016;351:aad3867.
60. Kervestin S, Jacobson A. NMD: a multifaceted response to premature translational termination. Nat Rev Mol Cell Biol. 2012;13:700–12.
61. Combier JP, de Billy F, Gamas P, Niebel A, Rivas S. Trans-regulation of the expression of the transcription factor MtHAP2-1 by a uORF controls root nodule development. Genes Dev. 2008;22:1549–59.
62. Hanyu-Nakamura K, Sonobe-Nojima H, Tanigawa A, Lasko P, Nakamura A. Drosophila Pgc protein inhibits P-TEFb recruitment to chromatin in primordial germ cells. Nature. 2008;451:730–3.
63. Pauli A, Norris ML, Valen E, Chew GL, Gagnon JA, Zimmerman S, Mitchell A, Ma J, Dubrulle J, Reyon D, et al. Toddler: an embryonic signal that promotes cell movement via Apelin receptors. Science. 2014;343:1248636.
64. Slavoff SA, Heo J, Budnik BA, Hanakahi LA, Saghatelian A. A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. J Biol Chem. 2014;289:10950–7.
65. Ferrin MA, Subramaniam AR. Kinetic modeling predicts a stimulatory role for ribosome collisions at elongation stall sites in bacteria. eLife. 2017;6:e23629.
66. Ivanov IP, Shin BS, Loughran G, Tzani I, Young-Baird SK, Cao C, Atkins JF, Dever TE. Polyamine control of translation elongation regulates start site selection on antizyme inhibitor mRNA via ribosome queuing. Mol Cell. 2018;70:254–264. e256.
67. Barbosa C, Peixeiro I, Romao L. Gene expression regulation by upstream open reading frames and human disease. PLoS Genet. 2013;9:e1003529.
68. Tarrant D, von der Haar T. Synonymous codons, ribosome speed, and eukaryotic gene expression regulation. Cell Mol Life Sci. 2014;71:4195–206.
69. Warner JR. The economics of ribosome biosynthesis in yeast. Trends Biochem Sci. 1999;24:437–40.
70. Duncan CD, Mata J. The translational landscape of fission-yeast meiosis and sporulation. Nat Struct Mol Biol. 2014;21: 641–7.
71. Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS. High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science. 2012;335:552–7.
72. Rodriguez CM, Chun SY, Mills RE, et al. Translation of upstream open reading frames in a model of neuronal differentiation. BMC Genomics. 2019;20:391.
73. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. Elife. 2014;3:e03523.
74. Hartenstein V, Campos-Ortega JA. Fate-mapping in wild-type Drosophila melanogaster. I. the spatio-temporal pattern of embryonic cell divisions. Roux Arch dev Biol. 1985;194:181–95.

75. Williams MJ. Drosophila hemopoiesis and cellular immunity. J Immunol. 2007;178:4711–6.
76. Pueyo JI, Magny EG, Sampson CJ, Amin U, Evans IR, Bishop SA, Couso JP. Hemotin, a regulator of phagocytosis encoded by a small ORF and conserved across metazoans. PLoS Biol. 2016;14:e1002395.
77. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. Landscape of transcription in human cells. Nature. 2012;489:101–8.
78. Heiman M, Kulicke R, Fenster RJ, Greengard P, Heintz N. Cell type-specific mRNA purification by translating ribosome affinity purification (TRAP). Nat Protoc. 2014;9:1282–91.
79. Chen X, Dickman D. Development of a tissue-specific ribosome profiling approach in Drosophila enables genome-wide evaluation of translational adaptations. PLoS Genet. 2017;13:e1007117.
80. Liang S, Bellato HM, Lorent J, Lupinacci FCS, Oertlin C, van Hoef V, Andrade VP, Roffe M, Masvidal L, Hajj GNM, Larsson O. Polysome-profiling in small tissue samples. Nucleic Acids Res. 2018;46:e3.
81. Patraquim P, Mumtaz MAS, Pueyo JI, Aspden JL, Couso, JP. Developmental regulation of Canonical and small ORF translation from mRNAs. GSE147619. 2020. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147619. Accessed 27 Mar 2020.

Publisher's Note